Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search

Peter Clark, John Thompson, Heather Holmback, Lisbeth Duncan
Mathematics and Computing Technology
The Boeing Company, PO Box 3707, Seattle, WA 98124
{peter.e.clark,john.a.thompson,heather.h.holmback,lisbeth.e.duncan}@boeing.com

Abstract

With the growth of on-line information, the need for better resource location services is growing rapidly. A popular goal is to conduct search in terms of concepts, rather than words; however, this approach is frequently thwarted by the high up-front cost of building an adequate ontology (conceptual vocabulary) in the first place. In this paper we describe a knowledge-based Expert Locator application (for identifying human experts relevant to a particular problem or interest), which addresses this issue by using a large, pre-built, technical thesaurus as an initial ontology, combined with simple AI techniques of search, subsumption computation, and language processing. The application has been deployed and in use in our local organization since June, 1999, and a second, larger application was deployed in March 2000. We present the Expert Locator and the AI techniques it uses, and then we evaluate and discuss the application. The significance of this work is that it demonstrates how years of work by library science in thesaurus-building can be leveraged using AI methods, to construct a practical resource location service in a short period of time.

Introduction

With the rapid growth of on-line information, it is becoming increasingly hard for users to find the information they need. The phenomenon of posing a query to a Web search engine and receiving many thousands of "hits", few of which are really relevant, is a familiar one. A well-known contributor to this problem is that search is organized around *words* (contained in the target documents) rather than the *concepts* which those words denote. As a word can denote many concepts (polysemy) and a concept can be denoted by many words (synonymy), a user's query may both miss relevant documents and hit irrelevant ones. In addition, without an unambiguous representation of what the user is interested in, it is impossible to apply domain knowledge to reason about the user's information request.

In this paper, we describe our recent work in conducting search in terms of *concepts* (unambiguous denotations of the entities of interest) rather than words, to reduce the ambiguity problem and also exploit domain knowledge for search. In particular, we have exploited an extensive technical thesaurus to provide both a conceptual vocabulary ("ontology") and a source of domain knowledge, avoiding the high up-front cost of ontology-building from scratch. We have combined this with simple AI techniques of search, subsumption computation, and language processing. We have used this for building an "Expert Locator" search tool for identifying human experts within our 200 person organization relevant to a user's problem or interest. This application has been deployed and in use within our organization since June 1999, and a similar, larger application was recently deployed in March, 2000, indexing a larger group of technical experts within Boeing. We describe the initial version of the thesaurus-based Expert Locator and the AI techniques that have been used to enhance it in various ways, and then we discuss and evaluate the application. Our conclusion is that, when the thesaurus and application domain are well matched, the many years of work by library science in thesaurus-building can be leveraged using AI methods to construct a practical resource location service.

Approach

A Thesaurus as a Conceptual Vocabulary

One challenge for working in concept space is the construction of an appropriate ontology ("conceptual vocabulary") appropriate to the domain of interest. To address this, we have used a technical thesaurus as the initial ontology, seeking to exploit the many years of effort already spent by librarians in constructing a conceptual vocabulary for a domain. Other alternative (but more costly) approaches would be to hand-build the ontology from scratch, e.g., CoalSORT (Monarch & Carbonell 1987), or learn it automatically from analysis of text corpora, e.g., PhraseFinder (Jing & Croft 1994).

It is important to note that a library thesaurus is distinct from a synonym dictionary (a common misconception) in two important ways. First, each term (concept) in the thesaurus has a unique name, precisely to remove word ambiguity. Sometimes concept names will include a parenthetical qualification if a single word would be ambiguous, e.g., "planes (geometry)", "beams (radiation)". Second, a thesaurus encodes not only the conceptual vocabulary but also semantic relationships (of a rather informal kind) between

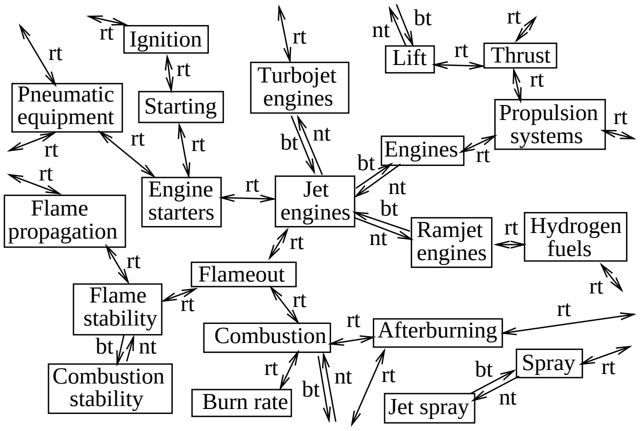

Figure 1: Sketch of a tiny fragment of Boeing's Thesaurus. The full Thesaurus contains approximately 37,000 concepts and 100,000 relationships between them.

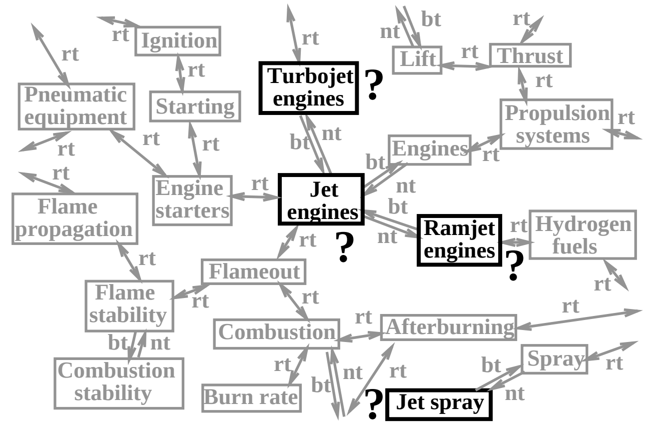

Figure 2: From the user's initial query word(s) (here the word "jet"), the system identifies possible concepts he/she may be referring to by simple stemming then substring matching.

concepts, the three most important types being named 'bt' (broader term), 'nt' (narrower term), and 'rt' (related to). A broader term denotes a subject area which encompasses the original term, usually[1] corresponding to a generalization (superclass) link in an inheritance hierarchy (e.g., "jet engines" –bt→ "engines"), while a narrower term is the inverse of this. The rt relation expresses that some (unspecified) close conceptual relationship exists between the two concepts. Although the semantics of these links are rather informal, they nevertheless provide (by design) knowledge about conceptual relationships in the domain, specifically for the task of information retrieval.

The particular thesaurus we have used is Boeing's Technical Thesaurus, built by Boeing Technical Libraries. This Thesaurus is a vast network of approximately 37,000 concepts (plus another 19,000 synonym concept names), with approximately 100,000 links between them (of the three types mentioned above), plus additional relationships ('subject note', 'used for', etc.) for other purposes. It is well suited to our purposes as it is highly customized to our target domain (aerospace) and organization (Boeing), and is rich in aerospace and "Boeing-speak" concepts, and also in concepts from the related areas in which Boeing is involved (e.g., computing, finance, sales, personnel management). A tiny fragment (0.05%) of this Thesaurus, sketched as a graph, is shown in Figure 1, where boxes denote Thesaurus concepts and arcs denote relationships.

Performing Concept-Based Search using a Thesaurus

For typical word-based search tools, an indexing engine (e.g., a Web crawler) builds ahead of time a word index of resources (e.g., Web pages) to be searched. At search time, a user enters a set of query words, and a matching algorithm then compares these with the word index to identify the "best" resources that match the user's query. Our goal is to do an analogous thing in concept space, requiring three main tasks to be performed:

1. A concept index needs to be built, in which target resources are indexed in terms of the concepts (not words) characterizing them.
2. A user's query needs to be (re-)formulated in terms of concepts.
3. A "concept-based search" algorithm is needed to match the user's concept query with the concept index of resources.

In our application, the Expert Locator, the "resources" we are interested in searching for are, in fact, not documents but human experts. We adopted straightforward approaches to the three tasks listed above:

1. The concept index was built manually, by asking each expert to characterize his/her area(s) of expertise by a list of concepts drawn from the Thesaurus.
2. After the user enters a set of search words, the system finds possible concepts he/she might be referring to (by stemming the user's words and then substring matching on concept names in the Thesaurus), and then asks the user to select his/her intended concept(s). This is illustrated in Figure 2, where the user has entered the word "jet", to which the system will ask: "By 'jet' did you mean: (i) jet engine (ii) ramjet engine (iii) jet spray (iv) ...?" If none of these are appropriate, the user can browse the Thesaurus by iteratively clicking on a concept to see its neighbors in the Thesaurus graph, to help locate his/her concept(s) of interest.
3. Thus having a set of concept(s) the user is interested in, the system searches for experts who either know about one of those concepts or know about concepts "closely related to" the user's concepts of interest, where "closely related to" corresponds to the distance between the user's

[1]but not always, for example "France" may be declared as a narrower term of "Europe", expressing a meronymic (part-of) rather than hypernymic (subclass) relation.

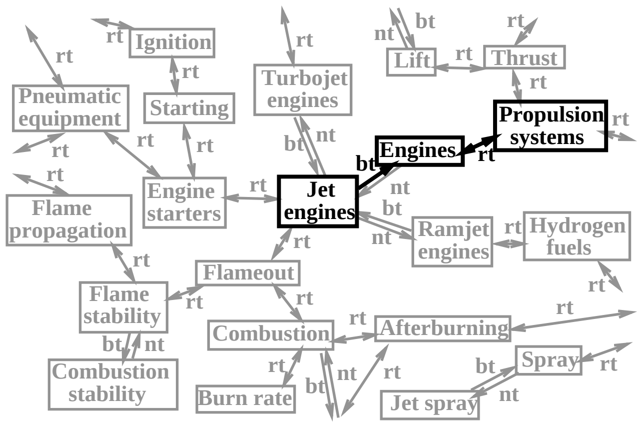

Figure 3: The distance between concepts in the Thesaurus graph is used to assess relevance between an expert's area of expertise (here "propulsion systems") and a user's concept of interest (here "jet engines").

and expert's concepts in the Thesaurus graph[2]. This is illustrated in Figure 3, where an expert in "propulsion systems" is considered moderately relevant to a query for information about "jet engines", as the concepts "propulsion systems" and "jet engines" are relatively close in the graph.

An example of the Expert Locator application itself is shown in Figure 4. The user asks for someone who knows about "graphics", to which the system asks (via a menu): "By 'graphics' did you mean: (i) graphic arts (ii) raster graphics (iii) graphic methods ...?" The user ticks the box(es) to identify his/her intended concept(s) (here "computer graphics"), then the system retrieves experts who either tagged themselves as having expertise in that concept(s) or in concepts closely related to it. In Figure 4, the system reports both the expert's technical area (e.g., "digital video") and the semantic path of associations from that concept to the user's original concept of interest (e.g., "via computer graphics → digital video").

Initial Results and Development

Initial, informal trials with the Expert Locator were encouraging, and in particular illustrated how even the semi-formal semantic relationships in the Thesaurus could help search. At the same time, these trials highlighted several issues which needed to be addressed:

1. The length of the shortest path between two concepts in the thesaurus graph is a crude measure of "semantic distance" or "relevance" between two concepts, and further tight controls on the graph search are needed.
2. Despite the 100,000 relationships in the Thesaurus, we frequently found cases where (at least for our purposes) desirable links were missing. In fact, of the 37,000 concepts in the Thesaurus, approximately 15,000 are "orphans" (i.e., not connected with any other concept), meaning that knowledge of concept associations could not be applied in those cases[3].
3. Even with 37,000 concepts, experts occasionally found that their area of expertise was missing, i.e., not properly characterized by any of the original Thesaurus concepts.

We now describe how these issues were (and continue to be) addressed.

Semantic Distance and Relevance

Evaluating semantic relatedness using network representations has a long history in artificial intelligence and psychology, e.g., (Resnik 1995; Brooks 1998; Tudhope & Taylor 1997), and we are exploiting this basic idea for our application. Intuitively, a short path (measured by counting links) between concepts in the Thesaurus graph might be expected to correspond to some loose notion of "relevance" between those concepts. For example, "artificial intelligence" and "semantics" are two links apart ("artificial intelligence" −rt→ "inference" −rt→ "semantics") and also intuitively appear to be related. Unfortunately, this is not always the case, as relevance is not always transitive. For example, "flood control" and "artificial intelligence" appear 'close' as they are just three links apart ("artificial intelligence" −rt→ "cybernetics" −rt→ "control" −nt→ "flood control"); similarly, "battle management" is within three links of "library science" ("battle management" −bt→ "management" −nt→ "information management" −rt→ "library science"). Table 1 gives a coarse quantification of this phenomenon using data from the evaluation (described later), showing the average relevance of concepts at distance D to an initial concept C. Due to this rapid degradation of relevance with distance, we severely constrain the search for relevant concepts to those at distance two or less from an initial concept. Work by other researchers suggests that this degradation might be partially reduced by weighting links using statistical methods, e.g., (Resnik 1998; Manning & Schutze 1999; Chen *et al.* 1993), an avenue which we are considering exploring.

One simple refinement we made to this "link counting" approach was to disallow paths which include a generalization (bt) link followed (either immediately or otherwise) by a specialization (nt) link. This avoids paths such as "battle management" −bt→ "management" −nt→ "information management", where the path first moves to a general concept and then specializes to a concept disjoint with the original. An informal evaluation of this approach suggests that this improves the credibility of the paths as explaining relevance (as these questionable paths are removed), but often alternative (more plausible) paths can be found and the list of relevant concepts is not significantly altered (e.g., blocking the path "battle management" −bt→ "management" −nt→ "information management" −rt→ "library science" causes

[2]This is an oversimplification; we elaborate shortly.

[3]One factor contributing to this relatively high proportion of orphans was that a separate database of "identifiers" (additional, non-Thesaurus keywords) was merged into the Thesaurus several years ago, with connections to Thesaurus terms still being added. As described later, our graph enhancement algorithm can assist with this linking process.

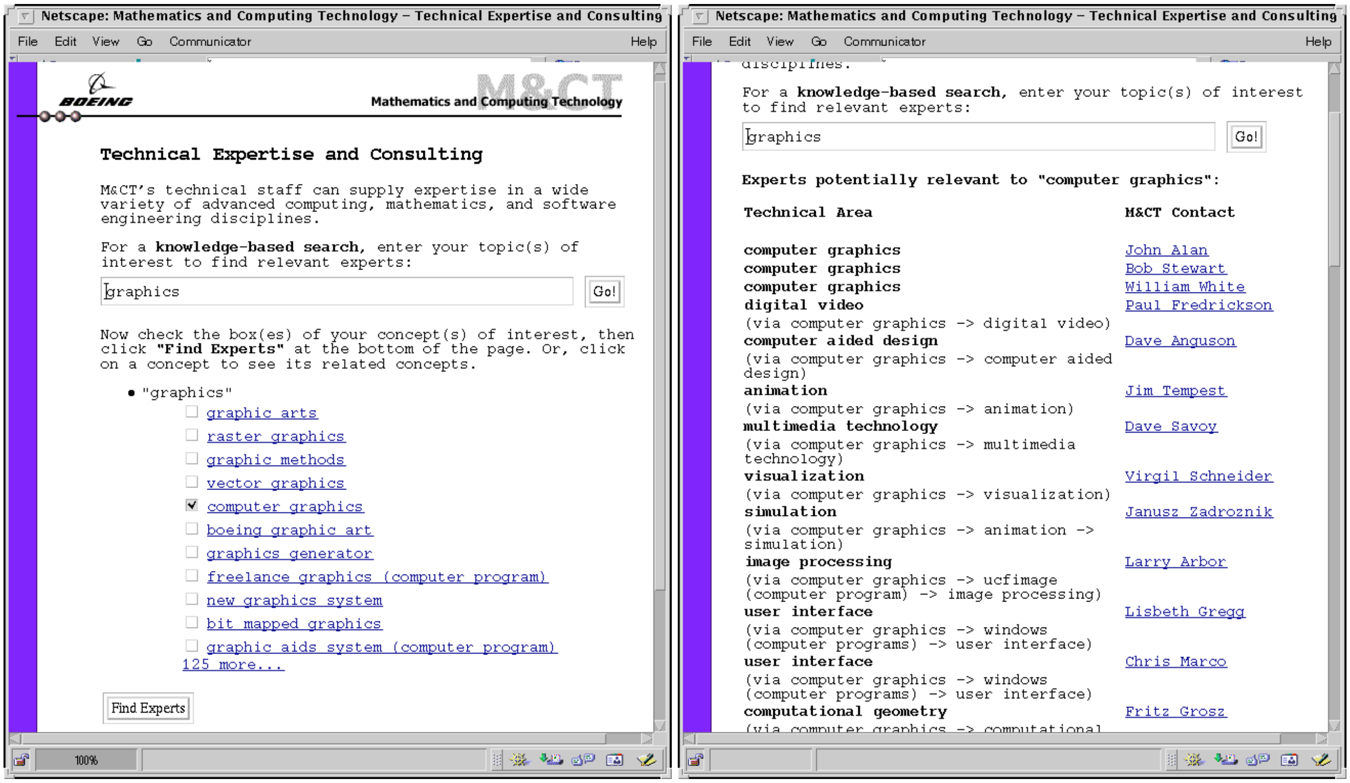

Figure 4: Two consecutive screen shots of the Expert Locator in use. First the user identifies the concept(s) he/she is interested in (by browsing and checking boxes), then experts are retrieved based on the proximity of their areas of expertise to those concept(s) in the semantic graph.

		Distance D from C			
		0	1	2	3
Number of concepts at D from C	(mean)	1	2.69	46	356
	(median)	1	1	10	96
Mean relevance of concepts at D to C		10.0 ± 0	7.8 ± 0.3	4.3 ± 0.3	3.2 ± 0.4

Table 1: Relevance degrades rapidly with distance, as judged by a human expert on a 0 (irrelevant) to 10 (relevant) scale. $\pm$ denotes standard error. The table also shows the rapid increase in concept accessibility with distance.

a different path "battle management" —rt→ "command control" —nt→ "information systems" —rt→ "library science" to be found instead).

Enhancing the Thesaurus Connectivity: Computing Extra Subsumption and Association Relationships

Although the Boeing Thesaurus is highly connected (100,000 links), it is often the case that desirable links, at least for our purposes, were missing, including 40% (15,000) of the 37,000 concepts being orphans and thus inaccessible to search.

However, an important characteristic of technical thesauri is that many concept names are compound (multi-word) terms. In Boeing's Thesaurus, 32,000 (85%) of the concept names are compound nouns or phrases. This allows some automated analysis of the concepts to be performed, based on the constituent words in these terms, using subsumption computation techniques (Woods 1991). For example, the concept "space shuttle main engine" is an orphan in the Thesaurus, but by comparing its constituents with other concept names, an algorithm can infer that it is related to the concept "space shuttle" (as "space shuttle" is a concept in the Thesaurus) and generalizes to "engines". Similarly, "metal pipe welding" can be inferred as a specialization of "tube joining", as "pipe" is a specialization of "tube" and "welding" is a specialization of "joining" in the Thesaurus.

We implemented a graph enhancement algorithm for this task, that automatically inferred these missing links using using word-spotting/natural language processing technology. This algorithm computes subsumption relationships between terms in a similar style to (Woods *et al.* 1999), and can be viewed as a simple classification engine using the limited semantics that a thesaurus affords. The algorithm

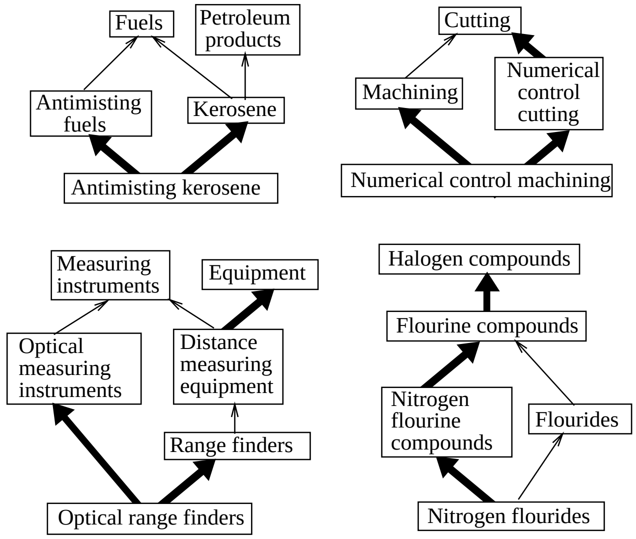

Figure 5: Four subgraphs of the enhanced Thesaurus, showing generalization (bt) links. The thin arcs were part of the original Thesaurus, the thick arcs were inferred automatically by the graph enhancement algorithm and added to the Thesaurus. "related-to" (rt) links (not shown) are inferred in a similar way. Approximately 58,000 links were inferred in total.

behaves as follows. First, individual words in a compound term are stemmed, and then the compound is generalized (in all possible ways) by repeatedly either removing the first word (e.g., "turbojet engine" becomes "engine") or generalizing one of its words/sub-phrases using the taxonomic links in the Thesaurus (e.g., "turbojet engine" becomes "jet engine"). If the generalization thus created exists in the Thesaurus as a concept in its own right, then taxonomic (bt/nt) links are added. This is repeated for all concepts, and then a final sweep of the resulting graph is performed to remove redundant taxonomic links (e.g., A−bt→C is considered redundant if A−bt→B and B−bt→C). This algorithm is extended to also add "related to" links by also allowing the last word in the compounds to be removed in the the 'generalization' step (e.g., "space shuttle engine" is related to "space shuttle"). Compounds created in this way (if they are also Thesaurus concepts), and any new ones derived from them, are then linked to the original concept using "related-to" (rt) links, rather than taxonomic (bt/nt) links.

Applying this algorithm, approximately 21,000 generalization/specialization links and 37,000 related-to links were automatically added, and the number of orphans was reduced from approximately 15,000 (≈40%) to 4,600 (≈13%). This approach relies heavily on the choice of name the Thesaurus authors decided to use for a concept. The algorithm will sometimes make mistakes due to changing word sense (e.g., "mean value analysis" becomes mistakenly related to the concept "values", in the sense of ethics), or finding unintended sub-phrases (mis-bracketing). However, interestingly, such mistakes were unusual, mainly attributable to the Boeing Thesaurus being a domain-specific rather than general thesaurus, where words in concept names tend to be used in a single sense (namely the aerospace sense). As a result, almost all of the (many) mis-generalizations hypothesized by the algorithm are non-Thesaurus concepts, and thus do not contribute links to the enhanced Thesaurus. We qualitatively evaluate the effect of the graph enhancement algorithm later in this paper.

Natural Language Processing of Compound Nouns

As a generalization of this approach, we have started work applying more sophisticated natural language processing technology to analyze compound nouns in the Thesaurus. This offers several advantages:

1. It provides better regularization of word variations (e.g., recognizing that "antisubmarine", "anti-submarine", and "anti submarine" are all variants of the same concept).
2. It can help disambiguate the appropriate word sense, based on the other words in the compound (e.g., "manual" in "transmission manual" refers to the concept of "manuals (documentation)", not "hand-operated").
3. It can help identify appropriate word grouping (bracketing), e.g., "advanced knowledge engineering" = "advanced (knowledge engineering)", not "(advanced knowledge) engineering".
4. It can refine the all-encompassing "related to" link into finer semantic categories, e.g., identifying that "coal" is the `result` of "coal mining", while "strip" is the `manner` of "strip mining".

An interactive prototype system called NCAS (Noun Compound Analysis System) was developed to perform this task. Word regularization, part of speech information, and identification of possible bracketings are performed by a standard parsing component. For word sense disambiguation, preferred bracketing, and identification of the head-modifier relation, we follow a knowledge-based linguistic approach of using a set of noun-noun (and also adjective-noun) interpretation rules. Similar approaches to noun compound interpretation have been performed by others, for example (Barker & Szpakowicz 1998; Vanderwende 1993; Finin 1986). An example of an interpretation rule is:

For compound "*modifier head*" (e.g., "metal tube"):
IF *modifier* is a `Material`
AND *head* is a `Physical-Object`
THEN *head* `is-made-of` *modifier*.

A set of 27 noun-noun relation types were chosen (e.g., `agent`, `causes`, `contains`, `location`) by merging relations from our previous NLP work (Holmback, Duncan, & Harrison 2000) with Barker's list (Barker & Szpakowicz 1998), and then these were augmented to fit the Thesaurus data. This latter step was based on manual analysis of the 450 most common noun compounds that were either a Thesaurus concept name or a subphrase of a concept name. As well as identifying the relation type, the rules constrain which word senses can co-occur. For example, the above rule constrains "tube" in "metal tube" to be a physical object (e.g., a pipe), thus ruling out "tube" in the senses of an abstract geometric shape or a subway.

Our work in this area is still preliminary, and the rule base, word sense classification hierarchy, and association of word senses with Thesaurus concepts are still incomplete. However, processing of noun compounds in this way, or more generally dictionary definitions, e.g., similar to MindNet (Dolan, Vanderwende, & Richardson 1993), may prove to be a useful additional way to augment the semantic knowledge base.

Handling non-Thesaurus Concepts

In addition to processing the concepts within the Thesaurus, the natural language processing of compound nouns offers a way of dealing with concepts that are missing in the Thesaurus, but are of interest to the user (either to express his/her area of expertise, or to perform a search). Currently, if an expert cannot find a suitable concept to characterize his/her expertise, a common strategy is to use a set of concepts, each representing an element of his/her desired concept name. For example, an expert in "document releasing" (missing in the Thesaurus) may tag him/herself with the concepts "documents" and "releasing". This is problematic because his/her expertise is not about releasing in general, but about a particular *type* of releasing (namely of documents). However, by tagging him/herself with the general "releasing" concept, he/she will be considered highly relevant to concepts neighboring "releasing" in the Thesaurus, such as "venting", "emission", etc., a clearly undesirable consequence.

Instead, we would like the system to accept this compound noun as a new concept represented as a structure denoting the relationship between its constituents, rather than as a set of independent concepts, which we can informally sketch as:

$$\text{"Document Releasing"} \neq \boxed{\text{Documents}} + \boxed{\text{Releasing}}$$

$$= \boxed{\begin{array}{l}\underline{\text{Releasing}} \\ \quad\text{object: Documents}\end{array}}$$

The noun-noun processing technology we have implemented can be used for exactly this task, by interactively (or automatically) linking the user's new concept to existing concepts, both for classifying resources and posing a search query. This would mark a significant shift in cataloging/classification from a task of concept *selection* to one of concept *construction* from primitives.

Evaluation

Finding good evaluation metrics for this style of application is challenging. The ultimate success of the Expert Locator application relies on several factors: the quality of the underlying knowledge-base (the enhanced thesaurus), the search algorithm, the ability of experts to label themselves appropriately with thesaurus concepts in the first place, the ability of users to identify their concepts of interest to perform a search, and other issues such as speed and the friendliness of the interface. Thus it is important to consider which aspect(s) of the system are being evaluated (the original Thesaurus? the enhancements? the experts' ability to describe themselves?). In addition, it is difficult to select what to compare the the Locator against (i.e., what constitutes "success"?).

There are several weak indicators of the system's utility that we can point to. The system is deployed and has achieved limited but sustained use (averaging approximately 1.2 searches per workday since June 1999), with 163 experts in our organization currently self-registered using 314 subject areas (concepts), mainly in the fields of computer science and mathematics (our organization's main technologies). Feedback has been very positive, and has spawned the construction of a second, larger application, indexing a separate, larger community of experts. This second application was deployed in March 2000, and has been used for 596 searches during its first three weeks of use (i.e., to time of writing), even though its availability has not been widely advertised yet. The most significant requirement people have pointed to is not with the concept-based search itself, but to restrict this search to a subset of the database constrained by simple attribute filters, e.g., years at Boeing, job type. This is a straightforward extension which we are planning to incorporate.

In a trivial way, the Expert Locator improves on simple word-based searches of an expertise database simply because, by definition, it does not require the user to enter exactly (or indeed any) of the subjects the experts classified themselves under, but will instead find "relevant" experts even if there is not an exact match with a user's query. Two specific questions are how the size of the search (i.e., the distance bound on the search from the initial concept(s), in number of links) affects precision and recall, and what effect the automatic enhancement of the original Thesaurus with subsumption and related-to links has had.

As a rough evaluation of this phenomenon, we performed an analysis in which a human expert selected a concept he knew about, and then scored a random sample of the concepts at distances 1, 2, and 3 away according to a subjective measure of "relevance", similar to semantic distance experiments in psychology, e.g., (Brooks 1998), on a score of 0 (completely irrelevant) to 10 (completely relevant). The assumption here is that to the extent a concept is relevant, an expert on that concept would be able to answer a question about the original selected concept. This assessment was performed using both the original Thesaurus alone and with the additional links automatically added.

We can use these measures to assess the Expert Locator's search as follows: for each concept C_i in the Thesaurus, let $r_{ij} = 0$ if concept C_j is deemed completely irrelevant to it, or 1 if it is deemed completely relevant. Thus, if a search for concepts relevant to a concept C_i retrieves concepts $C_1, \ldots, C_N$, then (using standard definitions) **precision** $= \sum_{j=1}^{N} r_{ij}/N$ (the proportion of hit concepts which are relevant), and **recall** $= \sum_{j=1}^{N} r_{ij} / \sum_{j=1}^{M} r_{ij}$ (the proportion of relevant concepts which are hit), where M is the total number of concepts in the Thesaurus. In our case, where we have 'degrees of relevance', we allow r_{ij} to also take fractional values between 0 and 1 (= (the manually judged relevance on the 0 to 10 scale)/10). As assessing $\sum_{j=1}^{M} r_{ij}$ (the total number of concepts relevant to C_i in the Thesaurus,

Radius D of search	Original Thesaurus Graph Precision (%)	Original Thesaurus Graph Relative Recall (%)	Enhanced Graph Precision (%)	Enhanced Graph Relative Recall (%)
0	100 ± 0	6 ± 2	100 ± 0	6 ± 2
1	84 ± 3	26 ± 6	80 ± 3	39 ± 1
2	58 ± 6	48 ± 7	57 ± 5	75 ± 9
3	50 ± 6	65 ± 10	42 ± 7	100 *

Table 2: Variation of precision and recall (relative to recall within distance 3 in the enhanced graph, *) in locating concepts relevant to some initial starting concept. $\pm$ denotes standard error. The results show that the Thesaurus enhancements significantly improve recall, with only a minimal negative effect on precision.

weighted by relevance) is impractical (M = 37,000), we instead assume all relevant concepts are within a distance three in the enhanced Thesaurus, and thus the recall scores are only relative to concepts in this set (hence "relative recall"). This assumption only affects the factor by which the recall scores are normalized, not their relative sizes, which is our main interest for this comparative study. The results, averaged over five different trials, i.e., for five different concepts C_i, are shown in Table 2.

These results suggest that enhancing the Thesaurus has had only a minimal negative effect on precision, while significantly increasing recall. In other words, the automatically added links are apparently of comparable quality, in denoting relevance, as the original manually added links, and allow a significantly larger number of relevant concepts (thus experts) to be identified during search. The occasional errors in the linking algorithm (e.g., due to not recognizing word sense change) is probably one contributing factor to the fractional difference in these figures.

Discussion, Critique, and Conclusion

Perhaps the most significant result of this work is to highlight the potential value of combining a technical thesaurus with simple AI techniques of search, subsumption computation, and language processing, allowing us to construct and deploy a practical expert location system in a very short time. Library science has spent many years building conceptual taxonomies in the form of thesauri, and the resources available there are sometimes overlooked in AI research. We have demonstrated how we can exploit this work for a practical task in combination with AI techniques, and have also speculated on more sophisticated AI methods which could be applied to further enhance the application.

In some ways, the utility of the Expert Locator is somewhat surprising, given the well-known difficulties in equating "number of links" with "relevance", e.g., (Resnik 1995). In fact, our experience largely confirms previous findings that, in general, link distance is a weak measure of relevance, and only in the restricted case of very short paths (lengths 1 or 2) was this a meaningful measure to use (Table 1), contrary to our initial expectations. A second point of note is that we are using a technical (rather than general) thesaurus, highly customized to our particular application domain and company's activities. This provides an important filter, as only aerospace/Boeing-specific concepts and relationships are present, thus automatically "biasing" the knowledge to just that required for the domain at hand. In fact our initial work started with WordNet (Miller *et al.* 1993) (a general-purpose lexical reference system of linked concepts), but was quickly abandoned precisely because many of the links it contained were irrelevant and detrimental to aerospace-specific queries. The domain-specificity of the Boeing Thesaurus not only constrains search by encoding just a domain-specific notion of relevance, but also constrains the Thesaurus enhancement algorithm to add only aerospace-relevant links (e.g., "giant hangar" will not be related to the concept "giants" precisely because the concept "giants" is not in the Thesaurus). In addition, mistakes from word ambiguity in concept names are significantly reduced, as words tend to be used in the same (aerospace) sense.

It is also clear that there are further developments which can be made. In particular, a list of concepts is a rather crude characterization of an expert's ability or a user's information need, and using structured representations would help considerably in this respect, as discussed earlier. Similarly, migrating the Thesaurus to a knowledge-base with more rigorous semantics would enable inferencing and question-answering services to be added, and provide a basis for computing relevance using more principled domain knowledge rather than concept associations. However, the simplicity of the presented approach is also a considerable strength – it has allowed a practical system to be built and deployed, and in a way which is easily reproducible by others. It also provides a springboard from which these refinements can now be explored.

Specific to the expert location task, we have assumed that expert relevance is equated with concept relevance. While there is obviously an important relationship, it is also clear there are other important factors which we have not taken into account, e.g., an expert's years of experience, location, and position in the company, which should be added for selecting the portion of the database to search. We have also not attempted to quantify the "quality" of an expert, e.g., through recommendations from others or "social filtering", as performed by so-called recommender systems (Kautz 1998) such as ReferralWeb (Kautz, Selman, & Shah 1997). This would be another possible dimension for expert location to explore.

Although we have focussed on expert location, there is essentially nothing in the presented approach which is specific to this task, and the same approach could be applied or integrated with search for other resource types, e.g., projects, documents, and work groups. Again, a concept-based index of the resource entities would be needed, which could be constructed either manually or (in the case of text) automatically using statistical methods (Manning & Schutze 1999). This points to the exciting possibility of using a thesaurus-derived knowledge-base for organizing and indexing a wide variety of information resources, again coupling many years work in library science with AI techniques to provide potentially valuable information management services, an avenue which we are currently exploring.

Acknowledgements:

We are greatly indebted to Boeing Technical Libraries, in particular Gail Shurgot, Mary Whittaker, and Corinne Campbell, for their work on the Boeing Thesaurus and their encouragement and support for this work. Thanks also to Don Retallack, Steve Woods, Mike Uschold, and Rob Jasper for ideas and feedback on this work and its application.

References

Barker, K., and Szpakowicz, S. 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. COLING-ACL'98*, 96–102.

Brooks, T. 1998. The semantic distance model of relevance assessment. In *Information Access in the Global Information Economy (Vol 35): Proceedings of the 61st Annual Meeting of ASIS*, 33–44.

Chen, H.; Lynch, K. J.; Basu, K.; and Ng, T. D. 1993. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert* 25–34.

Dolan, W. B.; Vanderwende, L.; and Richardson, S. 1993. Automatically deriving structured knowledge-bases from a machine-readable dictionary. In *Proc. of the Pacific Assoc. for Computational Linguistics*.

Finin, T. W. 1986. Constraining the interpretation of nominal compounds in a limited context. In Grishman, R., and Kittredge, R., eds., *Analyzing Language in Restricted Domains*. NJ: Erlbaum. 163–173.

Holmback, H.; Duncan, L.; and Harrison, P. 2000. A word sense checking application for simplified english. In *Third International Workshop on Controlled Language Applications*.

Jing, Y., and Croft, W. B. 1994. An association thesaurus for information retrieval. Technical Report UM-CS-1994-017, Dept CS, Univ Massachusetts at Amherst.

Kautz, H.; Selman, B.; and Shah, M. 1997. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM* 40(3).

Kautz, H., ed. 1998. *Proc. AAAI Workshop on Recommender Systems*. AAAI Tech Report WS-98-08.

Manning, C. D., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. MA: MIT Press.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. 1993. *Five Papers on WordNet*. NJ: Prinston Univ. (http://www.cogsci.princeton.edu/∼wn/).

Monarch, I., and Carbonell, J. 1987. CoalSORT: A knowledge-based interface. *IEEE Expert* 39–53.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In Mellish, C. S., ed., *IJCAI-95*, 448–453. Kaufmann.

Resnik, P. 1998. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of AI Research* 11:95–130.

Tudhope, D., and Taylor, C. 1997. Navigation via similarity: Automatic linking based on semantic closeness. *Information Processing and Management* 33(2):233–242.

Vanderwende, L. 1993. SENS: The system for evaluating noun sequences. In Jensen, K.; Heidorn, G. E.; and Richardson, S. D., eds., *Natural Language Processing: The PLNLP Approach*. Boston, MA: Kluwer.

Woods, W. A.; Bookman, L. A.; Houston, A.; Kuhns, R. J.; Martin, P.; and Green, S. 1999. Linguistic knowledge can improve information retrieval. Tech Report TR-99-81, Sun Labs. (http://www.sun.com/research).

Woods, W. A. 1991. Understanding subsumption and taxonomy: A framework for progress. In Sowa, J. F., ed., *Principles of Semantic Networks*. CA: Kaufmann. 45–94.